MILLIMAN WHITE PAPER

Machine learning approaches to outlier detection

Abdal Chaudhry
Miruna Dudceac
Michael Leitschkis

Milliman

Life insurance companies have to project thousands of simulations over 30 to 60 years due to the long-term nature of insurance liabilities and the complex guarantees embedded into life insurance products. In particular, Internal Model firms are required to produce a full risk distribution in their Solvency II work to capture risk interdependencies (cross-terms and impact of inter-risk correlations).

A brute-force nested simulations approach to this task is computationally infeasible with current resources available to insurers because it would require 1,000 or more risk-neutral simulations to produce a single stochastic scenario of the full risk distribution. A brute force approach would then require several thousand such points to work out a full loss distribution in an empirical manner. Because of this, several proxy modelling approaches have been developed and refined over the last 15 years—such as the so-called Least Squares Monte Carlo (LSMC) approach.

Under LSMC, we approximate a dependent variable such as net asset value (NAV) or best estimate liability (BEL) using a polynomial in numerous relevant risk drivers such as nominal yields, corporate bond spreads or longevity trend. In order to calibrate an LSMC polynomial, we have to generate several thousand training points, each of which represents a combined stress to all risk drivers with the total training data set covering the full calibration range. Note that the calibration range covers a wide spectrum of outcomes for each risk driver, because it has to cover upward and downward 1-in-200-year stresses and extreme stresses to all risk drivers.

The dependent variable for each training data point is generated by running combined risk driver stresses through the life insurer's fund cash flow model. Most actuarial models are not built keeping in mind extreme scenarios such as, e.g., extreme low interest rates and negative spreads, and therefore a cash flow model may produce unreasonable NAVs for a few severe combined stresses. These questionable values will then be among the inputs used by the LSMC proxy modelling calculation engine—unfortunately, even one unreasonable training point input can degrade the goodness-of-fit of the LSMC polynomial across the calibration range.

Because we cannot reasonably expect any manual inspection of thousands of training points by actuaries within any realistic working day timetable, our only hope is to implement some automated outlier removal and/or anomaly detection algorithm.

In this context, the aim of this paper is to find a way to identify and remove those points that qualify as outliers in the sense that:

(1) They are not plausible and may be due to cash flow model error.

(2) Their value is significantly impacting the model fitting.

By doing this there should be an improved and more stable LSMC fitting.

The remainder of our paper is organised as follows:

- In Section 2, we look into a very simple outlier deletion technique and realise its limitations.
- In Section 3, we discuss a powerful method known as *Cook's distance*.
- In Section 4, we consider a few alternative machine learning approaches.
- In Section 5, we draw conclusions and outline a few areas for further research.

While most of this paper addresses the LSMC use case, we hope that it gives users valuable insights into anomaly detection methods irrespective of whether they use LSMC or not.

DATA AND GOODNESS-OF-FIT MEASURES

The data used in this paper consists of 40,000 training data points. Each point within the data set consists of 35 different risk drivers representing market and insurance risks and a single dependent variable, say, the NAV.

Each risk driver value is based on Sobol numbers and takes values between -2 and 2. The independent variables are uniformly distributed and fill the 35-dimensional risk space as densely as possible.

The Tukey's criterion

One of the more common and simpler methods used for outlier deletion is the Tukey's method. Under this method, an outlier is defined as being the observation which falls outside a chosen interquartile range.

The formula for two given quartiles, Q1 and Q3, lower and upper quartile and a nonnegative constant k is:

$$[Q1 - k * (Q3 - Q1), Q3 + k * (Q3 - Q1)]$$

An observation is considered to be an outlier whenever it falls outside the calculated range. The name Tukey's comes from John Tukey, who proposed that k should be equal to 1.5 and anything that falls outside that interval should be considered as an outlier [2].

The main drawback of this method is that it is one-dimensional, as it only considers the dependent variable while paying no attention to explanatory variables or, in our case, risk drivers. Because of this, a simple Tukey's outlier detection method will tend to remove certain scenarios that appear extreme in their magnitude (compared to other scenarios) but are not anomalies, as the underlying stresses are extreme. In other words, the behaviour of the fund cash flow model might actually be as expected when factoring in the explanatory variables. Hence, deleting these outliers is counterproductive and leads to a loss of information in the extreme areas of the risk space, which are of utmost interest to an insurer.

The table in Figure 1 shows some key statistics such as the average relative error, R^2 and mean squared error (MSE) which were obtained when the simple Tukey's method was used to detect and remove outliers from our training data. Our fitted polynomial was obtained using a forward step algorithm with the maximum order of single terms of 4 and a maximum cross order of 4. Unless otherwise stated, this would be the case throughout the remainder of this paper.

FIGURE 1: KEY STATISTICS

THRESHOLD LEVEL	AVERAGE RELATIVE ERROR	R2	MSE
NO THRESHOLD	195.21%	0.9415	88M
K = 1.5 (1401 POINTS DELETED)	193.99%	0.9338	69M
K = 1.3(1940 POINTS DELETED)	198.72%	0.9302	67M

Note that in the following sections we will be using the same goodness-of-fit criteria to compare different outlier detection methods.

By looking at the key statistics it is clear that:

- Even though the average relative error has improved under the first threshold, R^2 has worsened
- Removing additional points by shrinking the threshold results in a worse fit

Thus, in a complex data frame, where extreme scenarios are run, using the Tukey's method to identify outliers is convenient from the process viewpoint, yet not ideal. It may, however, help in detecting erroneous model code, e.g., if certain cash flow model code "blows up" the outcome for dependent variables in certain scenarios.

Cook's distance

METHOD DESCRIPTION

One method that has proven itself to truly improve the fit of the model by detecting and deleting outliers is Cook's distance.

Cook's distance is widely used in regression analysis to find influential points or points that negatively impact the fitted model. Under this method, each training data point's Cook's distance is measured to identify points with the highest Cook's distances, which are then removed based on a predefined threshold.

High values for D_i are of particular interest, because they measure discrepancies in fitting by removing a single point, hence qualifying this single point as "important" and even "outlier" in some instances. These points create an instability in model prediction and represent a source of significant estimation error.

For a given training data point, its Cook's distance is given by:

$$\mathrm{Di} = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2}$$

where:

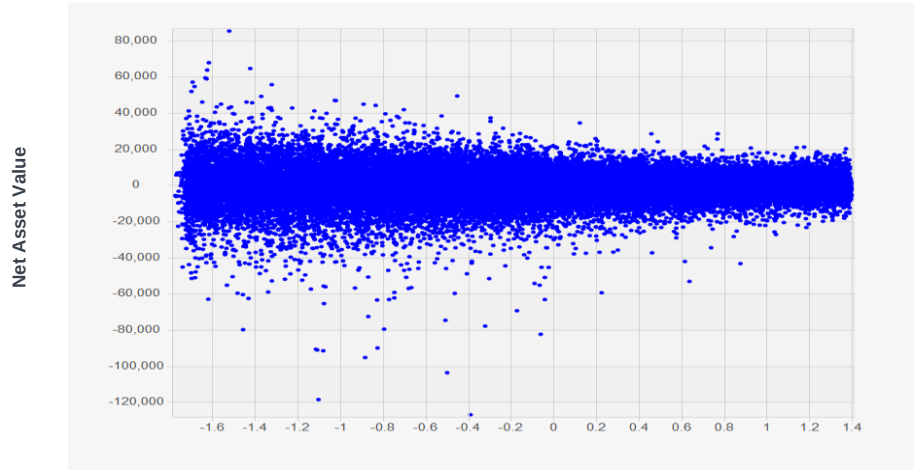
- $\mathrm{D_i}$ represents the Cook's distance for point i
- $\hat{y}_j$ represents the matrix of the fitted values
- $\hat{y}_{j(i)}$ represents the matrix of the fitted values when point i is deleted
- p represents the number of risk drivers
- s^2 is the mean squared error of the regression model

APPLICATION OF COOK'S DISTANCE TO REAL WORLD FITTING ISSUES

LSMC proxy modelling example

Figure 2 shows residuals obtained from the regression model fitted to the training data set consisting of 40,000 scenarios against the first risk driver in the data set.

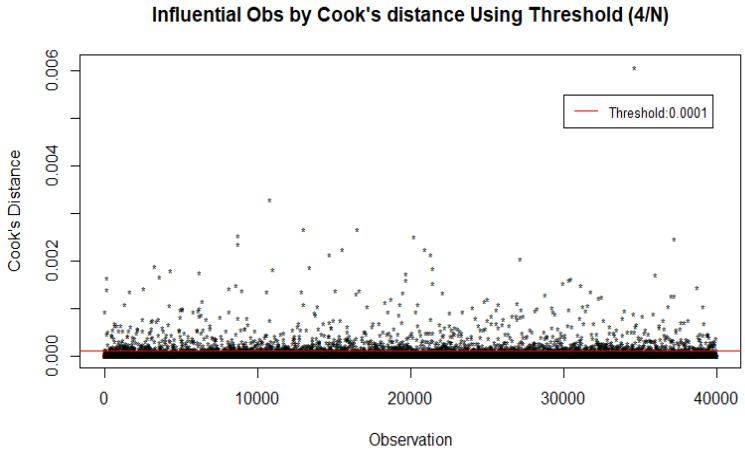
FIGURE 2: RESIDUALS OF INITIAL FIT



A clear residual funnel is observed, i.e., residuals are heteroscedastic. This is a problem in regression models and is discussed in detail in [1]. For a model fit to be considered adequate the variance of residuals should be uniformly distributed along the x-axis, and as it can be seen this is not the case here.

The aim of the Cook's distance approach is to try and identify points with high residuals and high leverage, i.e., high Cook's distances. Cook's distance is calculated for each observation and plotted in Figure 3 (x-axis here represents scenario ID).

FIGURE 3: COOK'S DISTANCE FOR EACH MODEL POINT



In this case, a threshold of 4/N has been used, where N represents the total number of points. This is one of the most commonly used thresholds for Cook's distance-based outlier deletion.

One of the more complicated aspects of using Cook's distance as an anomaly detection method is to pick an appropriate threshold. While 4/N is more commonly used in literature, one must consider an appropriate threshold to fit the purpose, as this will dictate how many points are removed. Picking a small threshold could lead to deletion of too many points and loss of relevant information while picking a larger threshold can lead to too few points being deleted resulting in a small change in the fitted model.

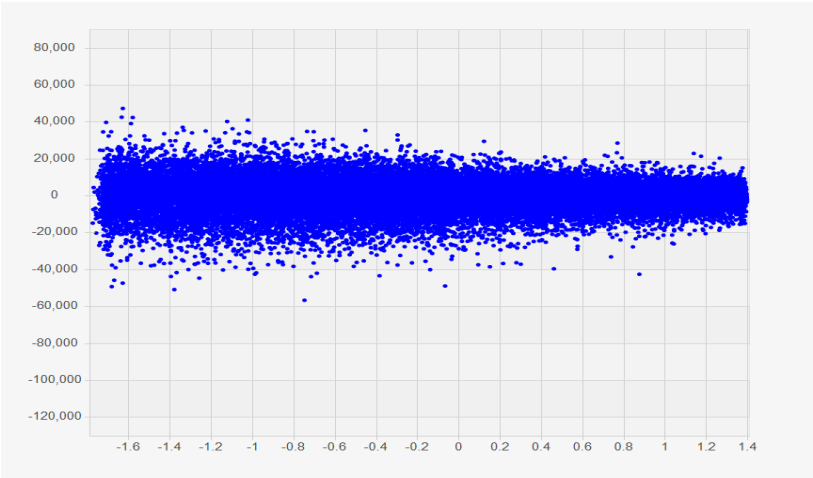
For this purpose, we consider three different thresholds:

1. The mean of the Cook's distance multiplied by 5.
2. Four divided by the total number of points (4/N).
3. The 45th quantile of the F-distribution fitted on the Cooks' distance points.

Plotting the results under the three thresholds we get the following results:

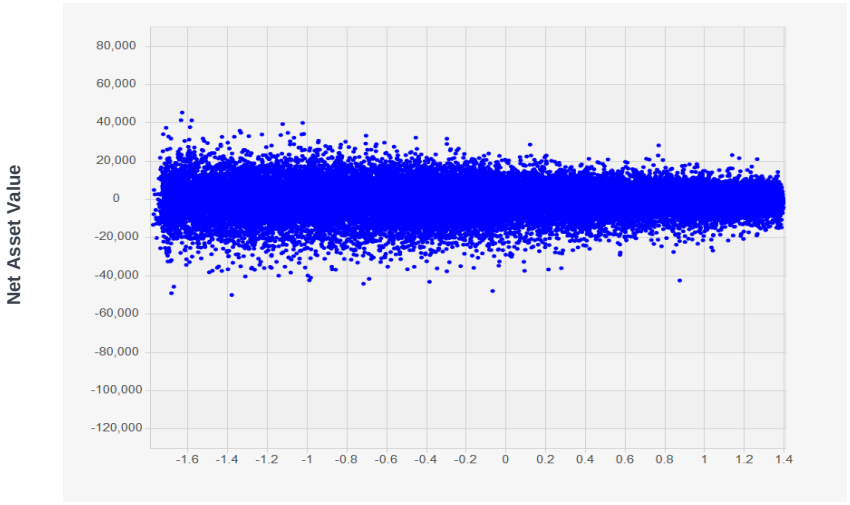
1. 5 x mean Cook's distance – deletion of 1,336 points

FIGURE 4: FITTED RESIDUALS AFTER OUTLIER DELETION FOR THRESHOLD OF 5X MEAN OF COOK'S DISTANCE



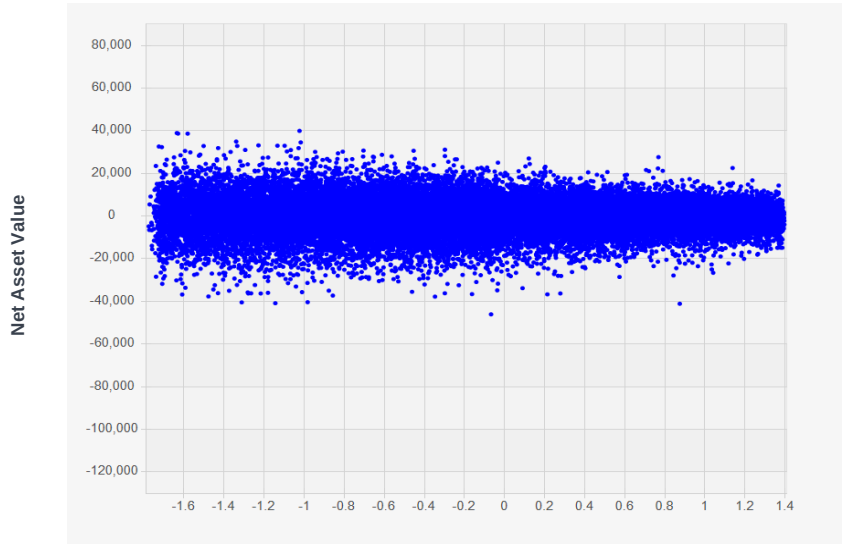
2. 4 / total number of points – deletion of 1,920 points

FIGURE 5: FITTED RESIDUALS AFTER OUTLIER DELETION FOR THRESHOLD OF 4/N



3. 0.45 quantile of the F-distribution – deletion of 2,637 points

FIGURE 6: FITTED RESIDUALS AFTER OUTLIER DELETION FOR THRESHOLD OF 0.45 F-DISTRIBUTION



The idea is that a cutoff point exists, after which deleting further points will start deteriorating the quality of the fit.

FIGURE 7: DECISION OF THE CUTOFF POINT FOR COOK'S DISTANCE

In the graph in Figure 7, the 95th percentile relative error under the four different thresholds—no outlier deletion, 5x mean Cook's distance, 4/N and the 45th quantile of the F-Distribution—decreases as more points are deleted. This means that the relative errors decrease for most model points as different thresholds lead to an increasing number of points being deleted more.

The figures above show that the fitted model improves significantly when points with high Cook's distances are removed. Furthermore, heteroscedasticity is visually improved. This is a direct consequence of using Cook's distance as an outlier deletion method, as points with high residuals and high leverage tend to have high Cook's distances. The results are clearer when looking at the in-sample goodness-of-fit statistics shown in the table in Figure 8.

FIGURE 8: GOODNESS-OF-FIT STATISTICS

THRESHOLD LEVEL	AVERAGE RELATIVE ERROR	R2	MSE
NO THRESHOLD	195.21%	0.9415	88M
5XAVERAGE COOK DISTANCE	177.65%	0.94677	61M
4/N	187.20%	0.94737	58M
QUANTILE OF F-DISTRIBUTION	198.61%	0.94875	54M

It is clear that as more points are removed the overall model improves—as at least two of the three test statistics improve for all threshold choices—but care must be taken not to remove too many points and lose relevant information that the data contains in those extreme scenarios. This is demonstrated by the increase in average relative error under the quantile of F-distribution approach, where the prediction power of the fitted model reduces for some extreme scenarios.

Alternative machine learning methods

In this section, we discuss a few alternative outlier detection methods that have been partially tested in an LSMC context as part of this paper. These range from clustering-based algorithms to Isolation forests. The remainder of this section will provide an overview and illustration of such methods that can be used as parallel or companion approaches with Cook's distance.

ANOMALY DETECTION USING CLUSTERING METHODS

The main purpose of a clustering-based outlier detection algorithm is to find clusters and outliers within these clusters. In this sense, a clustering-based outlier detection algorithm is a classification problem. Under this approach, there may be some outliers that are far away from other data points whereas other outliers would be nearby other data points. It makes sense that the outliers located far away will have more influence on the fit when it comes to LSMC—something that we have observed in our Cook's distance example above.

It is therefore advisable to look for outliers that are far away from other data points. In this section, we discuss an outlier detection approach using k-means clustering. The approach can be summarised as follows.

1. Choose, at random, the initial centroid for clusters.
2. Calculate the distance between cluster centroids to each training data point. In our example, we use Euclidean distance.
3. Assign each point to a centroid using minimum distance as the allocation criterion.
4. Repeat the steps 2 and 3 until the algorithm reaches stability.

Once clusters have been set, we need a measure to determine which points are our candidate outliers. For this purpose, we can use distances from the centroids calculated above and use the X^{th} *percentile* or the *mean* distance as our threshold. Points that sit outside this threshold are our candidate outliers and are taken to the next step. Note that, at this stage, we are only identifying our outlier candidates; we are not yet removing any data points.

The final stage involves setting outlier scores, or in other words picking our outliers from the candidate set described above. To compute the outlier score of a training data point *p*, the *k* nearest neighbours of it are found and:

1. The mean distance from *p* to all *k* nearest neighbours is calculated.
2. The average distance among the *k* nearest neighbours is calculated.
3. The ratio of *1 and 2* gives the outlier score.

Once the outlier score has been calculated, we remove *N* data points with the highest outlier scores. The algorithm can be run for a number of *N* and different number of clusters of the training data set. For the results presented in the table in Figure 9, we have chosen to remove approximately the same number of data points as were removed for Cook's distance. This ensures we fit the proxy model on the same number of training points to allow us to directly compare any improvements between k-means and Cook's distance.

FIGURE 9: RESULTS

# OF CLUSTERS	R2	MSE
BASE	0.94514	88M
25 CLUSTERS	0.9451	81M
100 CLUSTERS	0.9457	80M
250 CLUSTERS	0.9450	80M

Note that no significant improvement is achieved using a k-means-based outlier detection algorithm. This is partly due to the nature of our LSMC training data. More specifically, the training data used in the above example relies on Sobol coordinates for independent variables. Sobol numbers are, by definition, uniformly distributed within the risk driver range and therefore not appropriate for clustering, as the underlying risk calibrations remain *hidden*.

ISOLATION FOREST

Isolation forests (IFs) are similar to random forests and are built on decision trees. In IFs, randomly subsampled data is processed in a decision tree structure where features are selected at random. Samples which travel deeper into the tree are less likely to be outliers as they would generally require more cuts to isolate them. On the other hand, samples that end up in shorter branches are more likely to be outliers as it is easier for the tree to isolate them from other observations. The algorithm works as follows.

1. A random subsample of the training data set is selected and assigned to a binary tree.
2. Branching is performed by first selecting a random feature and random threshold.
3. Where the value of a point is less than the threshold it goes to the left branch. If not, it goes to the right branch.
4. Steps 2 and 3 are run recursively until either each point is completely isolated or until the maximum depth of the tree is reached.
5. Steps 2, 3 and 4 are repeated to construct random binary trees.
6. Once the group of trees is produced, the training is complete.

Once the training is complete, each training data point is traversed through all trees of the IF and an anomaly score is assigned to it. The anomaly score is an aggregation of the depth obtained from each of the trees. Data points identified as anomalies based on a user-defined threshold can then be removed.

This method is widely preferred in machine learning applications as it detects outliers purely based on isolation without having to employ a distance or density measure as in k-means above.

In an LSMC context, and in particular within our implementation of LSMC, IFs cannot be easily used as the risk driver information is *hidden* behind the Sobol coordinates used as risk driver values. However, our current work has led us to look at the use of IFs on LSMC data with multiple actuarial variables instead of a single dependent variable and on training data sets featuring the underlying risk calibrations that feed into Sobol coordinates.

Conclusion and outlook

Cook's distance outperforms other simple outlier detection methods when applied to LSMC proxy models. Our tests show that Cook's distance was able to improve the fitted regression model based on in-sample test results, reduction in heteroscedasticity of residuals and out-of-sample validation tests.

There are competing machine learning techniques that can lead to further improvements such as isolation forests and density-based spatial clustering algorithms. As we have seen, these techniques are relatively difficult to implement on the data set selected, compared to Cook's distance, and may not necessarily lead to significant improvements in the fitted models.

However, an advantage of these approaches is that they are agnostic to the predictive model whereas Cook's distance relies on the fitted model by measuring its sensitivity to the outliers removed. Therefore, we conclude that Cook's distance could be preferred as an outlier detection technique when validating proxy models developed for Solvency II Internal Model calculation purposes, while other methods discussed here could still produce valuable insights in other data science applications.

Literature

[1] Leitschkis M. & Horig M. (January 2012). Solvency II Proxy Modelling via Least Squares Monte Carlo. Milliman Research Report. Retrieved 14 December 2021 from https://ie.milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/life-published/solvency-ii-proxy-modelling.ashx.

[2] Chaudhry, A., Cherchali, A., Leitschkis, M., & Vedani, J. (August 2021). LSMC Surgery. Milliman White Paper. Retrieved 14 December 2021 from https://www.milliman.com/en/insight/lsmc-surgery.

[3] Tukey, J. (1977). *Exploratory Data Analysis*.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT
Abdal Chaudhry
abdal.chaudhry@milliman.com

Miruna Dudceac
miruna.dudceac@milliman.com

Michael Leitschkis
michael.leitschkis@milliman.com